

## OHIO TEST OF ENGLISH LANGUAGE ACQUISITION (OTELA) MARCH 2012 ADMINISTRATION

### STATISTICAL SUMMARY

These statistics describe the population of Ohio limited English proficient (LEP) students completing all domains in the March 2012 OTELA administration.

Test Grade Cluster/ Subject	N-count	Max Raw Score	Raw Score Mean	Raw Score Standard Deviation	Raw Score SEM	Max Scaled Score	Scaled Score Mean	Scaled Score Standard Deviation	Scaled Score SEM	Reliability
Grade K Listening	5338	21	12.62	5.26	1.29	432	335.16	54.28	13.30	0.94
Grade K Speaking	5338	24	14.99	6.42	1.29	429	329.30	59.36	11.90	0.96
Grade K Reading	5338	42	22.53	10.64	2.34	383	305.83	30.45	6.70	0.95
Grade K Writing	5338	27	12.92	6.67	1.64	391	282.32	42.77	10.53	0.94
Grade 1-2 Listening	9839	21	15.35	4.72	1.19	416	346.39	51.12	12.90	0.94
Grade 1-2 Speaking	9839	24	18.02	5.37	1.20	422	351.81	53.09	11.85	0.95
Grade 1-2 Reading	9839	42	28.01	9.97	1.95	396	322.89	38.95	7.62	0.96
Grade 1-2 Writing	9839	27	18.34	6.14	1.52	404	315.13	47.21	11.65	0.94
Grade 3-5 Listening	11360	18	12.09	3.56	1.78	928	708.68	129.38	64.72	0.75
Grade 3-5 Speaking	11360	24	20.04	4.49	1.63	929	793.97	129.30	46.93	0.87
Grade 3-5 Reading	11360	20	12.72	4.36	1.86	922	657.53	144.71	61.80	0.82
Grade 3-5 Writing	11360	17	9.52	3.39	1.86	954	625.49	130.70	71.73	0.70
Grade 6-8 Listening	6628	18	13.12	3.50	1.65	935	785.55	122.53	57.77	0.78
Grade 6-8 Speaking	6628	24	20.89	4.87	1.40	941	847.19	127.23	36.45	0.92
Grade 6-8 Reading	6628	20	12.60	4.40	1.86	926	675.60	142.64	60.38	0.82
Grade 6-8 Writing	6628	17	9.69	2.97	1.48	964	665.87	110.48	55.07	0.75
Grade 9-12 Listening	6104	20	13.46	4.01	1.87	942	781.68	119.93	55.88	0.78
Grade 9-12 Speaking	6104	24	20.45	5.42	1.40	949	858.08	132.69	34.31	0.93
Grade 9-12 Reading	6104	20	11.87	4.69	1.93	933	690.67	145.81	60.03	0.83
Grade 9-12 Writing	6104	20	11.96	4.06	1.88	995	708.25	135.48	62.83	0.78

**OTELA Cut Score Points for All Performance Standards**

		Performance Standard Cut Scores			
		Beginners	Intermediate	Advanced	Full English Proficiency
Grade K Listening	Raw Score	4	9	15	19
	Scaled Score	248	300	355	399
Grade K Speaking	Raw Score	6	12	18	22
	Scaled Score	255	300	349	394
Grade K Reading	Raw Score	8	20	36	40
	Scaled Score	270	300	338	359
Grade K Writing	Raw Score	7	16	21	26
	Scaled Score	251	300	328	375
Grade 1–2 Listening	Raw Score	6	11	16	19
	Scaled Score	254	300	348	382
Grade 1–2 Speaking	Raw Score	8	13	18	22
	Scaled Score	266	300	344	388
Grade 1–2 Reading	Raw Score	10	22	31	39
	Scaled Score	262	300	328	364
Grade 1–2 Writing	Raw Score	8	17	21	25
	Scaled Score	245	300	329	369
Grade 3–5 Listening	Raw Score	5	8	11	13
	Scaled Score	450	544	645	725
Grade 3–5 Speaking	Raw Score	5	10	17	22
	Scaled Score	450	547	668	809
Grade 3–5 Reading	Raw Score	6	11	13	17
	Scaled Score	450	580	648	770
Grade 3–5 Writing	Raw Score	5	9	12	14
	Scaled Score	450	577	669	785
Grade 6–8 Listening	Raw Score	7	9	12	14
	Scaled Score	554	626	718	806
Grade 6–8 Speaking	Raw Score	3	12	18	22
	Scaled Score	458	611	719	825

		<b>Performance Standard Cut Scores</b>			
		<b>Beginners</b>	<b>Intermediate</b>	<b>Advanced</b>	<b>Full English Proficiency</b>
Grade 6–8 Reading	Raw Score	6	11	14	18
	Scaled Score	460	612	690	829
Grade 6–8 Writing	Raw Score	7	10	12	15
	Scaled Score	553	653	722	894
Grade 9–12 Listening	Raw Score	6	9	12	16
	Scaled Score	556	632	729	850
Grade 9–12 Speaking	Raw Score	8	12	18	21
	Scaled Score	570	650	765	850
Grade 9–12 Reading	Raw Score	7	10	14	17
	Scaled Score	545	630	718	850
Grade 9–12 Writing	Raw Score	6	10	13	16
	Scaled Score	509	631	719	850

Note: Scale score cuts may not be observable on all forms and may not correspond directly to the attainable raw score in each category. Observable scale scores are presented in the raw to scale score conversion tables below.

**Percentage of Students at Each Performance Level**

Test Grade Cluster/ Subject	Percentage of Students at Each Performance Level				
	Pre-functional	Beginners	Intermediate	Advanced	Full English Proficiency
Grade K Listening	5.00	18.83	37.22	23.51	15.44
Grade K Speaking	9.84	18.83	31.96	22.12	17.25
Grade K Reading	9.39	30.89	46.46	7.79	5.47
Grade K Writing	19.65	43.82	22.35	11.43	2.75
Grade K Comprehension	9.31	30.67	46.78	7.85	5.40
Grade K Production	18.49	41.85	25.67	11.30	2.70
Grade K Composite	18.53	42.66	28.36	8.36	2.10
Grade 1-2 Listening	3.59	12.96	29.18	23.19	31.08
Grade 1-2 Speaking	4.80	9.99	25.61	27.55	32.05
Grade 1-2 Reading	5.22	20.40	28.68	29.31	16.38
Grade 1-2 Writing	6.71	25.69	24.02	28.78	14.80
Grade 1-2 Comprehension	5.15	20.02	29.43	29.35	16.04
Grade 1-2 Production	6.46	24.22	26.11	29.03	14.18
Grade 1-2 Composite	7.38	25.86	29.40	27.32	10.04
Grade 3-5 Listening	2.74	9.31	18.84	18.62	50.49
Grade 3-5 Speaking	1.58	2.28	12.30	35.32	48.53
Grade 3-5 Reading	7.29	23.04	13.18	34.04	22.46
Grade 3-5 Writing	7.84	30.18	31.55	17.75	12.68
Grade 3-5 Comprehension	5.92	20.52	17.98	33.80	21.79
Grade 3-5 Production	4.74	22.93	41.96	17.98	12.39
Grade 3-5 Composite	7.38	23.76	34.61	26.97	7.28
Grade 6-8 Listening	6.25	5.45	15.74	17.71	54.86
Grade 6-8 Speaking	1.45	5.30	8.84	18.53	65.89
Grade 6-8 Reading	7.53	24.23	20.20	34.31	13.73
Grade 6-8 Writing	15.63	25.92	28.46	27.76	2.23
Grade 6-8 Comprehension	6.14	20.40	25.69	34.19	13.58

<b>Test Grade Cluster/ Subject</b>	<b>Percentage of Students at Each Performance Level</b>				
	<b>Pre-functional</b>	<b>Beginners</b>	<b>Intermediate</b>	<b>Advanced</b>	<b>Full English Proficiency</b>
Grade 6-8 Production	9.35	17.50	43.23	27.70	2.22
Grade 6-8 Composite	10.49	21.15	39.54	27.50	1.31
Grade 9-12 Listening	4.39	9.65	14.45	34.24	37.27
Grade 9-12 Speaking	5.54	3.23	9.91	10.99	70.33
Grade 9-12 Reading	16.45	16.81	24.12	23.38	19.25
Grade 9-12 Writing	8.45	16.45	23.66	32.21	19.23
Grade 9-12 Comprehension	12.47	19.13	26.05	23.38	18.97
Grade 9-12 Production	6.49	11.99	30.26	32.50	18.76
Grade 9-12 Composite	11.17	19.48	29.51	30.67	9.17

## **Equating and Scaling: How Raw Scores Are Converted into Scaled Scores**

### **Test Form Construction**

The Ohio Test of English Language Acquisition (OTELA) is based on the English Language Development Assessment (ELDA) developed under the direction of a consortium of 18 member states of the LEP State Collaborative on Assessment and Student Standards (LEP-SCASS) and the Council of Chief State School Officers. The ELDA was designed to allow states to meet federal requirements under NCLB concerning the annual assessment of LEP students regarding their acquisition of and progress toward developing English language proficiency in listening, speaking, reading, and writing.

The OTELA is a battery of tests designed to allow schools to measure progress in the acquisition of English language proficiency skills among non-native English-speaking students. The battery consists of separate tests for listening, speaking, reading, and writing, for each of five grade clusters: K, 1–2, 3–5, 6–8 and 9–12. The tests are aligned with Ohio’s English language proficiency standards and were constructed to provide content coverage across four academic topic areas (English Language Arts; Mathematics, Science and Technology; and Social Studies), and one non-academic topic area, School-Environmental, which is related to aspects of the school environment such as extracurricular activities, student health, homework, classroom management, and lunchtime. Although the OTELA tests measure language skills with content drawn from age-appropriate curricular and non-curricular sources, they are not tests of academic content. Students do not need any external or prior content-related knowledge to respond to the test questions.

To measure a wide range of English language proficiency, the full-length ELDA includes many items and requires substantial test administration time. Although administration of the ELDA test battery is not officially timed, general guidelines indicate approximately four hours of test administration time. In addition, most students to whom the ELDA was administered scored in the upper ranges of the raw score distribution. These performance results indicated that the ELDA operational forms could be shortened substantially by eliminating the easiest items in the operational item bank while maintaining a proportional representation of items across content standards within each domain.

OTELA items were selected on the basis of their psychometric properties, contribution to measurement at key points on the scale (such as the intermediate cut score) and content coverage. When, for example, the easiest items within a domain proved to be concentrated within specific content standards, the Ohio Department of Education (ODE) opted to maintain breadth of content coverage, rather than to simply increase form difficulty. In addition, although a primary goal was to reduce test length as much as possible, estimated form reliabilities were used to determine the appropriate number of items to include in each test form.

### **Common Item Equating**

*Grade clusters 3–5, 6–8, and 9–12.* Following the first operational administration of grades 3–12 ELDA forms in 2005, items included in the first operational test forms were recalibrated, with the resulting item parameter estimates serving as the reference scales for ELDA. All subsequent grades 3–12 ELDA test forms are linked to these scales.

Because the first set of operational forms were constructed to include a set of common items between adjacent grade clusters, the grades 3–5, 6–8 and 9–12 forms were jointly calibrated in a single Winsteps run for each domain, resulting in a common, vertically linked scale across grade clusters for each domain. For each Winsteps run, the mean of the item difficulty parameters was fixed to zero so that the average difficulty for all items across grade clusters was equal to zero within each domain for the first operational form.

For the 2005 field test, a common item design was used to allow common item equating across field-test forms and the first operational form. Following the common item design of the field test, items were jointly calibrated in a single Winsteps run for each domain and grade-cluster combination. Because all of the 2005 ELDA field-test forms shared items in common with operational Form 1, a common item equating method was used to link the field-test items to the ELDA operational Form 1 scale. For each field-test form within each grade cluster, shared items were fixed to their operational Form 1 parameter estimates, while the remaining items were freely estimated. This placed all the field-test items on the operational Form 1 scale.

In addition, a small subset of items were field tested in 2004 but were not included in the 2005 operational forms. These items were also placed on the 2005 operational ELDA scale. Because all items in the 2005 operational test came from the 2004 field-test item pool, the 2005 operational test items were used as linking items. The mean-mean procedure was used to find the linking constant. To ensure that the final set of anchor items (i.e., common items) was free of item parameter drift, a stepwise deletion procedure was used to select anchor items and calculate the linking constant needed to bring the field test items onto the reference scale defined by the first operational administration. Following this procedure, a linking constant was calculated, using all anchor items, and then applied the linking constant to bring the items back to the reference scale. Anchor item parameter estimates were then examined to determine whether the difference between any adjusted or linked parameter estimates and the reference scale parameter estimates was greater than .3 logits. At each step, the item with the greatest difference between its linked and reference item parameter estimates was eliminated from the anchor set, provided the difference was greater than .3. A new linking constant was then computed and applied to the test items and the parameter estimates for the remaining anchor items were again examined to determine whether any exceeded the .3 tolerance level. This process was repeated until all remaining anchor items met the tolerance-level specifications. The linking constant was computed on the basis of this final anchor item set, and then applied to the 2004 ELDA field-test item parameters.

The result of these analyses was to place all items in each of the grade 3–12 ELDA domain item banks on the common scale defined by the first operational administration.

Additional items were subsequently developed for the OTELA assessment program and these items were embedded in the operational test forms for the 2009 and 2010 administrations of OTELA. Operational and embedded field test items were concurrently calibrated. The operational test items were used to link items from the 2009 and 2010 operational administration to the original ELDA scale. The average item difficulty for the operational test items were then computed based on both the spring 2009 and 2010 operational administration and the bank item parameter estimates from the original ELDA operational administration to identify the linking constant necessary to bring the 2009 and 2010 operational item parameters back to the ELDA reference scale. The resulting linking constant was then applied to the field test items to place the field test item parameter estimates on the original ELDA scale.

*Grade clusters K and 1–2.* Items in the grades K and 1–2 OTELA forms were calibrated independently of the items in the grades 3–12 scales and are not reported on the vertical scale used to report scores on the grades 3–12 OTELA tests. A large proportion of items in the listening and speaking tests are common across the grades K and 1–2 test forms, while item overlap between the grades K and 1–2 reading and writing test forms is minimal. Consistent with this perspective, item difficulties for the kindergarten and grades 1–2 OTELA test forms were calibrated following two distinct strategies. Parameters for all OTELA kindergarten and grade 1–2 items were estimated using Masters’ partial credit model, an extension of the Rasch model for polytomous items. Student item scores were obtained from the Spring 2006 operational administration of the OTELA. For the reading and writing assessments, items in each of the grades K and 1–2 operational test forms were calibrated in separate Winsteps runs. For the listening and speaking items, parameters for items in both the grades K and 1–2 forms were estimated simultaneously in a joint calibration. Once the listening and speaking items were calibrated, the resulting cross-grade item parameter estimates were used to generate form-specific raw score to theta scale conversion tables.

Reporting scales for the grades K and 1–2 OTELA forms were established by setting the “intermediate” or level 3, performance standard for each of the assessments to 300. Therefore, for both the grades K and 1–2 assessments, and across the four English language domains assessed, a score of 300 indicates attainment of an intermediate level of English language proficiency. The standard deviation of the scale was set to 15.

### **Performance Standards**

The OTELA is designed to provide student performance-level assessment results that are fully comparable with those from the ELDA. To achieve this goal, the OTELA uses the same performance standards adopted by the LEP-SCASS for the ELDA. Performance levels range from Full English Proficiency, a level at which an LEP student is deemed to be able to function effectively and consistently through the medium of academic English in the school system (and thus ceases to be defined as LEP), to Pre-functional, a level at which an LEP student is consistently unable to communicate with any success in the English of the school environment, although the student may have some limited knowledge of English. Student performance levels are reported for each of the four language domain scores, as well as for English language comprehension (derived from student performance on the listening and reading tests), production (derived from student performance on the speaking and writing assessments), and a composite performance level that reflects student performance in both English language comprehension and production.

#### **OTELA Performance Levels**

<b>Level</b>	<b>Label</b>
5	Full English Proficiency
4	Advanced
3	Intermediate
2	Beginners
1	Pre-functional

In the process of adopting ELDA performance standards for the OTELA, ODE, in consultation with the Ohio LEP Advisory Committee, elected to revise one ELDA performance level cut score. In the ELDA performance standards for writing, students in the grade 3–5 cluster must substantially outperform students in both the 6–8 and 9–12 grade clusters to achieve Full English Proficiency. To address this issue, a linear regression approach was used to identify a cut score for Full English Proficiency at the grades 3–5 cluster from the cut scores identified for Beginning, Intermediate, and Advanced performance levels on the grade 3–5 writing assessment. This analysis identified a cut score of 2.08 (in the theta metric; 867 on the ELDA reporting scale) for the Full English Proficiency cut score at the 3–5 grade cluster. AIR submitted the cut score and estimated impact data for the revised performance standard to the Ohio LEP Advisory Committee for their consideration. The Ohio LEP Advisory Committee recommended that ODE adopt the revised performance standard, which ODE has done.

While performance levels for the four domain tests (Listening, Speaking, Reading, and Writing) are based on scaled scores, performance levels for the three derived scores (Comprehension, Production and Composite) are based on the performance levels of the underlying domain tests. The Comprehension performance level is based on the set of rules relating student performance levels on the Listening and Reading domain tests shown in the table below. Following these rules, if a student performed at level 3 on the Reading test and at level 2 on the Listening test, then the student would receive a level 3 for English language Comprehension. If the levels were reversed, so that a student performed at level 2 on Reading and level 3 on Listening, then the assigned Comprehension performance level would be 2.

### Rules for Combining Listening and Reading Levels to Yield Student Comprehension Level

Rules for Combining Listening and Reading Levels to Yield Student Comprehension Level		
If <i>Reading</i> Level is:	And <i>Listening</i> Level is:	Then <i>Comprehension</i> Level is:
1	1	1
	2	1
	3	1
	4	2
	5	2
2	1	2
	2	2
	3	2
	4	2
	5	3
3	1	2
	2	3

<b>Rules for Combining Listening and Reading Levels to Yield Student Comprehension Level</b>		
<i>If Reading Level is:</i>	<i>And Listening Level is:</i>	<i>Then Comprehension Level is:</i>
	3	3
	4	3
	5	3
4	1	3
	2	3
	3	4
	4	4
	5	4
5	1	3
	2	3
	3	4
	4	5
	5	5

Similarly, performance levels for Production are based on the set of rules shown below describing the relationship between Speaking and Writing performance levels. For example, a student performing at level 5 on the Writing test and at level 4 on the Speaking test would receive a 5 for English language Production. If the levels were reversed, however, so that the student performed at level 4 in Writing and level 5 on the Speaking test, then the Production performance level would be set to 4.

**Rules for Combining Writing and Speaking Levels to Yield Student Production Level**

<b>Rules for Combining Writing and Speaking Levels to Yield Student Production Level</b>		
<i>Writing Level is:</i>	<i>And Speaking Level is:</i>	<i>Then Production Level is:</i>
1	1	1
	2	1
	3	1
	4	2
	5	2

<b>Rules for Combining Writing and Speaking Levels to Yield Student Production Level</b>		
<i>Writing Level is:</i>	<i>And Speaking Level is:</i>	<i>Then Production Level is:</i>
2	1	2
	2	2
	3	2
	4	2
	5	3
3	1	2
	2	3
	3	3
	4	3
	5	3
4	1	3
	2	3
	3	4
	4	4
	5	4
5	1	3
	2	3
	3	4
	4	5
	5	5

Performance levels for Comprehension and Production are in turn evaluated to create an overall Composite level using the rules shown below. When the Comprehension and Production performance levels are not the same, the rule is to average the two levels and round down. For example, if the performance level for Production were 3 and the performance level for Comprehension were 4, the average would be 3.5, and the final Composite performance level would be reported as 3.

**Rules for Combining Comprehension and Production Levels to Yield Student Composite Level**

<b>Rules for Combining Comprehension and Production Levels to Yield Student Composite Level</b>		
<i>If Production Level is:</i>	<i>And Comprehension Level is:</i>	<i>Then Composite Level is:</i>
1	1	1
	2	1
	3	2
	4	2
	5	3
2	1	1
	2	2
	3	2
	4	3
	5	3
3	1	2
	2	2
	3	3
	4	3
	5	4
4	1	2
	2	3
	3	3
	4	4
	5	4
5	1	3
	2	3
	3	4
	4	4
	5	5

**Spring 2012 Raw Score to Scaled Score Conversion Table—Grades K–2**

Raw Score	Scaled Scores Corresponding to Raw Score Points							
	Grade K Listening	Grade K Speaking	Grade K Reading	Grade K Writing	Grade 1–2 Listening	Grade 1–2 Speaking	Grade 1–2 Reading	Grade 1–2 Writing
0	195	187	213	175	178	183	198	167
1	210	202	226	190	193	198	210	183
2	225	216	240	205	208	213	222	198
3	237	227	248	216	220	224	230	208
4	248	236	254	226	231	233	236	217
5	260	246	259	235	242	242	241	224
6	272	255	263	243	254	250	246	231
7	282	264	267	251	264	258	250	238
8	291	272	270	257	273	266	254	245
9	300	279	273	263	282	273	258	251
10	309	286	276	269	291	280	262	258
11	318	293	279	275	300	286	265	264
12	327	300	282	280	309	293	269	270
13	336	307	284	285	319	300	272	276
14	346	315	287	290	328	308	275	282
15	355	323	289	295	338	316	279	288
16	365	331	291	300	348	324	282	294
17	375	340	294	305	358	334	285	300
18	386	349	296	310	369	344	288	307
19	399	359	298	316	382	354	291	314
20	416	369	300	322	399	364	294	321
21	432	380	302	328	416	376	297	329
22		394	304	334		388	300	337
23		411	306	341		405	303	345
24		429	308	349		422	306	356
25			311	359			309	369
26			313	375			312	386
27			315	391			315	404
28			317				318	
29			319				321	

Scaled Scores Corresponding to Raw Score Points								
Raw Score	Grade K Listening	Grade K Speaking	Grade K Reading	Grade K Writing	Grade 1–2 Listening	Grade 1–2 Speaking	Grade 1–2 Reading	Grade 1–2 Writing
30			321				325	
31			324				328	
32			326				332	
33			329				335	
34			332				339	
35			335				343	
36			338				348	
37			342				352	
38			346				358	
39			352				364	
40			359				372	
41			371				384	
42			383				396	

**Spring 2012 Raw Score to Scaled Score Conversion Table—Grades 3–12**

Raw Score	Scaled Scores Corresponding to Raw Score Points											
	Grade 3–5 Listening	Grade 3–5 Speaking	Grade 3–5 Reading	Grade 3–5 Writing	Grade 6–8 Listening	Grade 6–8 Speaking	Grade 6–8 Reading	Grade 6–8 Writing	Grade 9–12 Listening	Grade 9–12 Speaking	Grade 9–12 Reading	Grade 9–12 Writing
0	161	205	149	214	191	301	153	222	207	284	162	241
1	236	276	184	286	284	362	207	301	300	353	251	317
2	329	347	278	359	376	422	302	380	392	423	345	393
3	388	390	338	404	435	459	362	432	451	465	404	440
4	434	423	383	441	480	486	407	472	496	498	450	476
5	473	450	421	473	519	509	445	507	534	524	487	506
6	507	473	454	503	553	528	478	540	567	548	520	533
7	539	494	485	533	585	546	509	571	598	569	551	559
8	570	514	514	563	615	563	537	602	626	589	579	584
9	599	533	541	594	644	579	565	634	654	608	606	609
10	629	552	568	627	674	594	592	668	680	626	633	635
11	660	570	595	662	704	609	618	704	707	644	660	662
12	691	588	623	702	736	624	646	745	734	662	687	692
13	726	607	651	750	770	639	674	792	763	679	715	724
14	764	625	682	811	808	655	704	852	792	697	745	760
15	810	644	715	894	853	671	737	903	825	715	777	803
16	869	664	752	924	907	689	774	933	862	734	814	857
17	914	685	797	954	921	708	819	964	906	754	859	903
18	928	708	856		935	729	877		915	775	898	936
19		733	905			753	908		928	798	916	980
20		761	922			782	926		942	825	933	995
21		795				818				857		
22		840				866				900		
23		904				915				924		
24		929				941				949		